Association Analysis

- Given a set of transactions, find rules that will predict the occurrence of an item based on the occurrence of other items in the transaction.
- Called sometimes Market Basket Analysis.

Examples

- On Thursdays, grocery store consumers often purchase diaper and beer together.
- Customer who purchase maintenance agreements are very likely to purchase large appliances.
- When a new hardware store opens, one of the most commonly sold item is toilet ring.

These three examples illustrate the three common types of rules produced by market basket analysis: the *useful*, the *trivial*, and the *inexplicable*.

- Once the pattern is found, it is often hard to justify. Even if justified, the results would not be useful.
  - The useful rule contains high quality, actionable information.

    The first example provides and justifiable(high quality), and actionable information.
  - Trival results are already known by anyone at all familiar with the business.

    In fact, we already know that customers purchase maintenance agreements and large appliances at the same time.
  - Inexplicable results seems to have no explanation and do not suggest a course of action.

    It is doubtful that further analysis of just the market basket data can give a credible explanation.
- Trivial and Inexplicable Rules occur most often

How Does Market Basket Analysis Work

- Market Basket analysis start with transactions containing one or more products or service offering and some rudimentary information about the transaction.
- Each of these transaction gives us information about which products are purchased with which other products.

Grocery store transactions

Customer	items
1	orange juice, soda
2	milk, orange juice, window cleaner
3	orange juice, detergent
4	orange juice, detergent, soda
5	window cleaner, soda

Co-occurrence Table

	OJ	Window Cleaner	Milk	Soda	Detergent
OJ	4	1	1	2	1
Window Cleaner	1	2	1	1	0
Milk	1	1	1	0	0
Soda	2	1	0	3	1
Detergent	1	0	0	1	2

- Orange juice and soda are more likely to be purchased together than any other two item.
- Detergent is never purchased with window cleaner or milk.
- Milk is never purchased with soda or detergent.

- These simple observations are examples of associations and may suggest a formal rule like:

  "If a customer purchases soda, then the customer also purchases milk."

- The question is how good is the rule?
- Support and Confidence.
  - Two of the five transactions include both soda and orange juice.

    These two transactions support the rule. i.e, the support for the rule is two out of five or 40%.

  - Every transaction that contains soda also contains orange juice, the rule has a confidence 100%.
  - However, the inverse rule, "if orange juice than soda," among 4 transactions with orange juice, only two also have soda. Thus, its confidence is 50%.

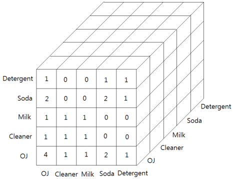
If A then B

Formally the support and the confidence are defined as follow:

- Support is defined as $\Pr(A \cap \mathrm{B})$ where the probability is the proportion of transactions contain both A and B among all transactions.
- While, confidence is defined as $\Pr(A \cap \mathrm{B}) / \Pr(A) = \Pr(\mathrm{B}|A)$

The Problem of Big Data

- The ideas behind the co-occurrence table extend to any combinations with any number of items, not just pairs of items.
- e.g, For combinations of three items, imagine a cube with each side split into five different parts.
- The 3-D co-occurrence table may produce rules such as "if A and B then C" or "if A then B and C."
- However, the number of combinations of a given size tends to grow exponentially.



- Suppose that a fast-food restaurant offers several dozen items on its menu, say there are a 100.

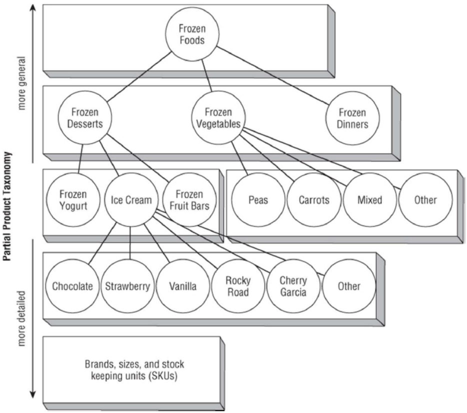
The number of combinations of items

# in combination	# of combinations
1	100
2	4,950
3	161,700
4	3,921,225
5	75,287,520
6	1,192,052,400
7	16,007,560,800
8	186,087,894,300

Choosing the Right Set of Items

- A grocery store may have tens of thousands of products on the shelves.
- Suppose a frozen pizza might be considered an item for analysis purposes—regardless of its toppings, its crust, or its size.
- So, the purchase of a large whole wheat vegetarian pizza contains the same frozen pizza.
- On the other hand, the manager of frozen foods may be very interested in the particular combinations of toppings that are ordered.
- Choosing the right level of detail is a critical consideration for the analysis.
- In the real world, items have product codes and stock-keeping unit codes that fall into hierarchical categories, called taxonomy.
- What level of taxonomy is the right one to use?

Taxonomies start with the most general and move to increasing detail



Generating Rules

- A rule has two parts, a condition and a result. *If condition, then result.*

  If 3-way calling, then call-waiting

- In practice, the most actionable rules have just one item as the result.
  “If diapers and Thursday, then beer”, is more useful than “If Thursday, then diapers and beer.”
- Thus, with 3 items A, B and C, the only rules to consider are:
  - If A and B, then C.
  - If A and C, then B.
  - If B and C, then A.

Probabilities of three items and their combinations

Combination	Probability
A	45%
B	42.5%
C	40%
A and B	25%
A and C	20%
B and C	15%
A and B and C	5%

Confidence in Rules

Rule	P(condition)	P(condition and result)	Confidence
If A and B then C	25%	5%	0.20
If A and C then B	20%	5%	0.25
If B and C then A	15%	5%	0.33

- The most confidence rule is the best rule, so we are tempted to choose "If B and C, then A."
- But there is a problem. The rule is actually worse than if just randomly saying that A appears in the transaction.
- A occurs in 45% of the transactions but the rule only gives 33% confidence.
- This suggests another measure called improvement.
- Improvement tells how much better a rule is at predicting the result than just assuming the result in the first place.

$$\text{improvement} = \frac{P(\text{condition and result})}{P(\text{condition})P(\text{result})} = \frac{\text{confidence}}{P(\text{result})}$$

- The improvement is similar to the Lift.

Improvement Measurement

Rule	Support	Condition	Improvement
If A and B then C	5%	20%	0.50
If A and C then B	5%	25%	0.59
If B and C then A	5%	33%	0.74
If A then B	25%	59%	1.31

- When improvement is greater than 1. then the resulting rule is better at predicting the result than random chance.
- None of the rules with three items shows any improvement.
- The rule “If A then B” is the best rule in the data.

Negating Rules

- When improvement is less than 1, *negating* the result produces a better rule.

```
If the rule
    "If B and C, than A"
has a confidence of 0.33, then the rule
    "If B and C, then, NOT A"
has a confidence of 0.67, and the new rule has an
improvement of 1.33.
```

- When the improvement score is low, it can increase them by negating the rules.
- However, negating rule may not be useful.

Minimum Support Pruning

- As the number of items in the combinations are growing, it requires heavy computation works.
- Thus, when then number of items are large, we need *pruning*.
- The most common pruning method employed in the association analysis is *minimum support pruning*. That is, a rule should hold on a minimum number of transactions.
- e.g., if there are 1 millon transactions and the minimum support is 1 percent, then only rules supported by 10,000 transactions are of interest.
- i.e., minimum support pruning eliminates items that do not appear in enough transactions.
- There are two ways to do this.
    - Eliminate the items from consideration.
    - Use the taxonomy to generalize the items so the resulting generalized items meet the threshold criterion.

Other Applications

- Baskets = documents
- Items = words in those documents
  - Find word that apper together unusually frequently, i.e., linked concepts.

	Word 1	Word 2	Word 3	Word 4
Doc 1	1	0	1	1
Doc 2	0	0	1	1
Doc 3	1	1	1	0

$$\text{Work 4} \Longrightarrow \text{Word 3}$$

Other Applications

- Baskets = sentences
- Items = documents containing those sentences
  - Sentences that appear together too often could represent plagiarism,

	Doc 1	Doc 2	Doc 3	Doc 4
Sent 1	1	0	1	1
Sent 2	0	0	1	1
Sent 3	1	1	1	0

$$\text{Doc } 3 \Longrightarrow \text{Doc } 4$$

Other Applications

- Baskets = Web pages
- Items = linked pages
  - Pairs of pages with many common references may be about the same topic.
- Baskets = Web pages p_i
- Items = pages that link to p_i
  - Pages with many of the same links may be mirrors or about the same topic.

	wp a	wp b	wp c	wp d
wp1				
wp2				

Reference

- Data Mining Techniques: For Marketing, Sales, and Customer Support, Michales J. A. Berry and Goron Linoff